Using context prior information in traffic sign detection with region proposals

Domen Tabernik, Alan Lukežič, Rok Mandeljc and Danijel Skočaj

Faculty of Computer and Information Science, University of Ljubljana, Slovenia
E-mail: {domen.tabernik,alan.lukezic,rok.mandeljc,danijel.skocaj}@fri.uni-lj.si

Abstract

In this paper we explore spatial and temporal context to augment the detection of traffic signs with a non-visual prior information. We propose to integrate the contextual information into a two-stage detector by re-scoring region proposals in the first stage. With the spatial prior, we explore how the frequency of traffic sign positions in images affects the recall rate, while with the temporal context, we capture the information from positions of known traffic signs obtained from the road maintenance authorities. We propose to project a 3D world location of each known traffic sign position into the 2D image position to prioritize specific region proposals. We show that the temporal context improves the recall rate for a two-stage detector on a database of 500 images.

1 Introduction

Traffic sign detection and classification is an integral part in road maintenance applications. Our previous work [1, 2] focused on the detection and recognition of traffic signs using region proposals designed specifically for road maintenance services; however, the specifics of this problem domain provide an additional prior knowledge that can be used to further improve the detector's performance. In this work, we explore how the prior knowledge, extracted from the non-visual contextual information, can improve traffic sign detection.

The addressed domain has two types of context information suitable for integration into the visual detector. One context information stems from the visual object domain itself. Traffic signs are always placed in locations that are in accordance with specific rules set by road maintenance authorities. Such rules make traffic signs appear frequently on the right side of the road, and at specific heights; on the other hand, very few traffic signs appear on the road itself or in the sky. Furthermore, in most cases road maintenance authorities have a record of all existing traffic signs and their exact positions in 3D world coordinates. During the detection stage, if we can supplement each image with approximate 3D world coordinates of traffic signs, then we can use this prior knowledge to reduce the search space for detection of the known traffic signs.

In this paper, we evaluate the integration of context information into a two-stage detector presented in our previous work [1]. We propose to integrate the context to improve the ranking of region proposals. We explore two types of non-visual prior information. Specifically, we use the spatial and the temporal context as a non-visual prior knowledge. We model both types of contexts as the likelihood of finding a traffic sign at specific image location. Contexts differ in the source of information they use to construct the likelihood. Spatial context uses the prior information that is specific for the domain of traffic signs, but is independent of the specific traffic sign instance or class. We collect the prior knowledge from frequencies of traffic sign occurrences in images. Temporal context, on the other hand, uses prior information obtained from the specific world location of known traffic sign instances. We project the 3D world location into the image location and assign a likelihood to each pixel. We explore spatial and temporal context on a dataset containing more than 500 images, and show that temporal context in particular improves the ranking of region proposals, thus leading to a better recall in object detection.

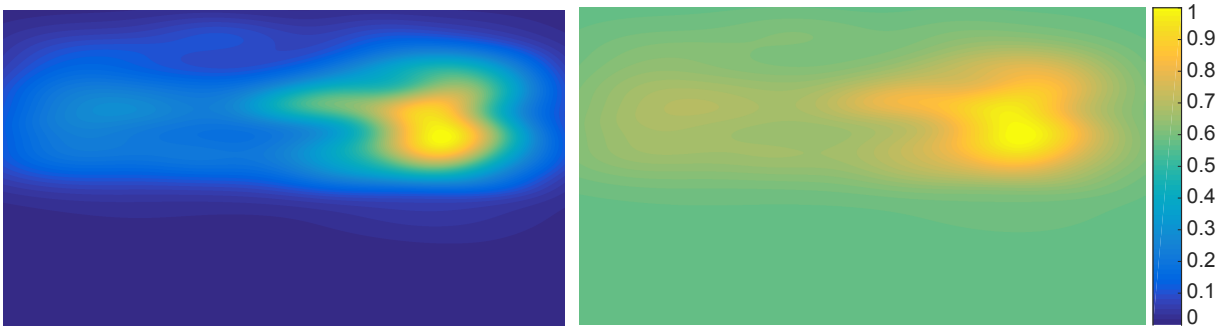
The paper is structured as follows: in Section 2, we present two types of contexts used, and evaluate them in Section 3. We conclude with the discussion in Section 4.

2 Contextual information in region proposals

In this section, we present the integration of contexts into a two-stage detector. We first provide a brief description of the two-stage detector, relevant for context integration, and refer the reader to [1] for a more detailed description.

The first stage of the detector consists of a region proposal algorithm, which generates a ranked list of regions that may contain a traffic sign. The second stage then classifies only N top-ranked regions with a feature classifier to obtain the exact class of the traffic sign. The best performance is obtained when all regions covering the traffic signs appear at the top of the list, preferably within $N = 1000$ or $N = 5000$ top-ranked regions out of all 100.000 regions. In the first stage, the ranked list of regions is generated using the edge-boxes [3] with the domain-specific learned structured edges [4], and an additional linear SVM. The final rank of a region r is defined as:

$$S(r) = S_{edge}(r) \cdot S_{svm}(r), \tag{1}$$



(a) (b)

Figure 1: Learned kernel density estimation model on ObcineDFG dataset (left) and the corresponding likelihood map with added uniform distribution using mixing factor $\alpha = 0.9$ (right). For better visualization, both maps are scaled to range 0–1.

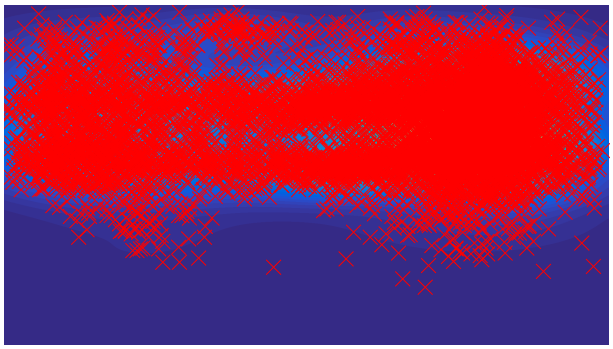


Figure 2: Positions of traffic signs (center points of annotations) on ObcineDFG dataset.

where $S_{edge}(r)$ is the score from edge-boxes, and $S_{svm}(r)$ is the score from linear SVM. Only the top 100.000 regions returned by the edge-boxes, and ranked by $S_{edge}(r)$, are re-scored using Eq. (1). The remaining regions are discarded.

We integrate context into region proposals by introducing an additional re-scoring factor, $S_{context}(r)$, with the goal of pushing the correct regions to the top of the list:

$$S(r) = S_{edge}(r) \cdot S_{svm}(r) \cdot S_{context}(r). \quad (2)$$

This is applied to the top 100.000 regions returned by the edge-boxes. We compute the context-scoring factor $S_{context}(r)$ as the likelihood of finding a traffic sign t at the center position of the region r:

$$S_{context}(r) = P(t|c(r)), \quad (3)$$

where $c(r)$ is the center position of the region r. We can compute the likelihood P in two ways, using two different contexts, i.e., the spatial and the temporal context. We present both types in the following subsections.

2.1 Spatial context

In spatial context, we capture the frequency of traffic sign positions in images. In Figure 2, we depict the distribution of center points of all regions containing traffic signs, as collected from more than 3.000 images of rural and city locations. Figure 2 demonstrates a distinctive pattern of traffic sign locations; they appear predominantly at the top of the image, and most frequently at the far left and far right positions.

We convert the frequency of region positions into a likelihood map by creating a kernel density map. We utilize multivariate online kernel density estimation with Gaussian kernels [5] to create the kernel density map. Furthermore, we smooth kernel density by increasing the variance of Gaussian kernels. We add an additional uniform component to prevent any potential traffic signs, appearing in the background, from being eliminated in the re-scoring process. The likelihood of a traffic sign appearing at the position $c(r)$, computed from the spatial context, is:

$$P_{sc}(t|c(r)) = (1 - \alpha) \cdot p(c(r)) + \alpha \cdot \mathcal{U} \quad (4)$$

$$p(x) = \sum_i w_i \Phi_{\Sigma_s}(x - x_i), \quad (5)$$

where $p(x)$ is the estimated kernel density using Gaussian kernel $\Phi_{\Sigma_s}(x - x_i)$ at position x, $\mathcal{U}$ is the uniform distribution, and $\alpha = 0.9$ is the mixing factor.

2.2 Temporal context

The temporal context encapsulates information about any previously-known 3D world position of traffic signs. This information may be obtained directly from road maintenance authorities, or from recordings of the previous runs along the same route.

To obtain the likelihood map from the known 3D world positions of traffic signs, we first project positions from the 3D world coordinates to the 2D image coordinates. An accurate GPS and depth data is required to achieve high precision. However, in our case, only X and Y axis of each traffic sign was estimated from the depth map, while we approximated the Z axis (i.e., height) from the known camera position and the annotated center of the traffic sign in the image. Such projected positions may contain a significant amount of noise due to calibration

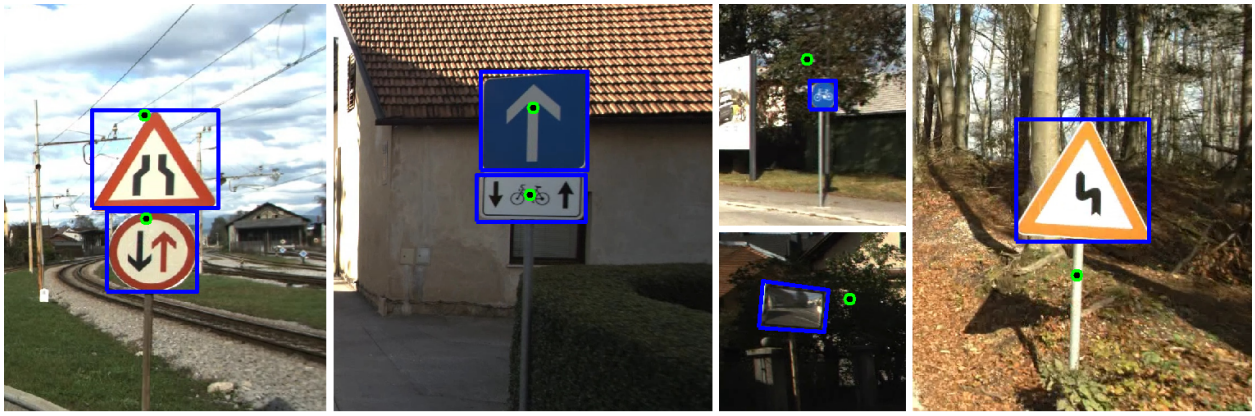


Figure 3: Example of several traffic signs from LjubljanaDFG2015 dataset. Blue rectangles show ground-truth annotations, while green circles are projected locations from LjubljanaDFG2014 dataset. The projection error can be severe in some cases (on the right) due to inaccuracies in calibration and poor GPS localization.

mismatches and inaccuracies in GPS positions (see Figure 3); we account for this noise by modeling the likelihood with a Gaussian kernel:

$$P_{tc}(t|c(r)) = \sum_i \Phi_{\Sigma_t}(c(r) - \tilde{p_i}), \qquad (6)$$

where $\Phi_{\Sigma_t}(c(r) - \tilde{p_i})$ is a Gaussian kernel centered at the projected point $\tilde{p_i}$. The variance of Gaussian is relative to image size. We use $1/4$ of the image size, however, in our case changing the variance did not influence the final results. If multiple traffic signs are projected into an image, the contributions of their Gaussian kernels are summed together.

3 Evaluation

We evaluate both contexts by measuring their effect on the performance of region proposals. A standard protocol to evaluate the performance of region proposals is to measure only the recall, while ignoring the precision. We measure the recall rate with respect to the allowed minimal intersection with the ground-truth regions, similar to the evaluation in [1]. As intersection-over-union measure we use the standard PASCAL VOC overlap [6].

3.1 Datasets

The evaluation was performed on a dataset with slightly over 500 images containing roughly 500 annotated traffic signs from both urban and rural areas.

For training, we use two additional datasets. The first one, named ObcineDFG and containing more than 3.000 images, was used to learn the spatial context. This dataset contains images from several cities and rural areas not captured in the testing set. The spatial context model, trained on ObcineDFG, is depicted in Figure 1a. This dataset was also used to train the linear SVM for region re-scoring, based on the edge information and using hard-negative mining from the background regions.

The second dataset is used to learn the temporal context. It contains the 3D world locations of the traffic signs that are captured and annotated in the testing dataset. These locations were obtained from the depth map and corresponding annotated images that were taken more than a year earlier under different visual conditions. The dataset also provides all necessary camera parameters for projecting the 3D world coordinates from training images into the 2D image coordinates of testing images.

3.2 Results

Results of the evaluation with and without context are shown in Figure 4. Improvements in recall are noticeable mostly at the high quality regions, i.e., at the higher intersection-over-union (IoU) overlap. Without the context, the top 5.000 regions cover only 62% of objects at 0.8 IoU and 97% at 0.5 IoU. The blue dashed line shows the best case for this region proposal algorithm, which captures 86% of objects at 0.8 IoU and 99.64% at 0.5 IoU, however, this requires running the second-stage classifier on all 100.000 regions.

Looking at the results with context, we can see that context helps to capture the remaining objects using only 5.000 top regions. However, the improvement is only noticeable with the temporal context. Specifically, the temporal context improves recall at the higher quality regions, while at 0.5 IoU, the improvement is only by 1–2%. At the higher quality regions, temporal context captures 75% of object at 0.8 IoU.

On the other hand, the spatial context does not appear to improve the recall rate. Changing parameters for the uniform distribution does not appear to improve the performance. The recall rate is even reduced when a lower mixing factor α is used for the uniform distribution. This may be an indication that the spatial context increases the rank of spurious regions. The spatial likelihood map prioritizes regions at specific locations, but in some images, those areas, while containing many region proposals, may not contain any traffic signs at all. The rank of such region proposals will be incorrectly increased, resulting in suppression of the correct regions.

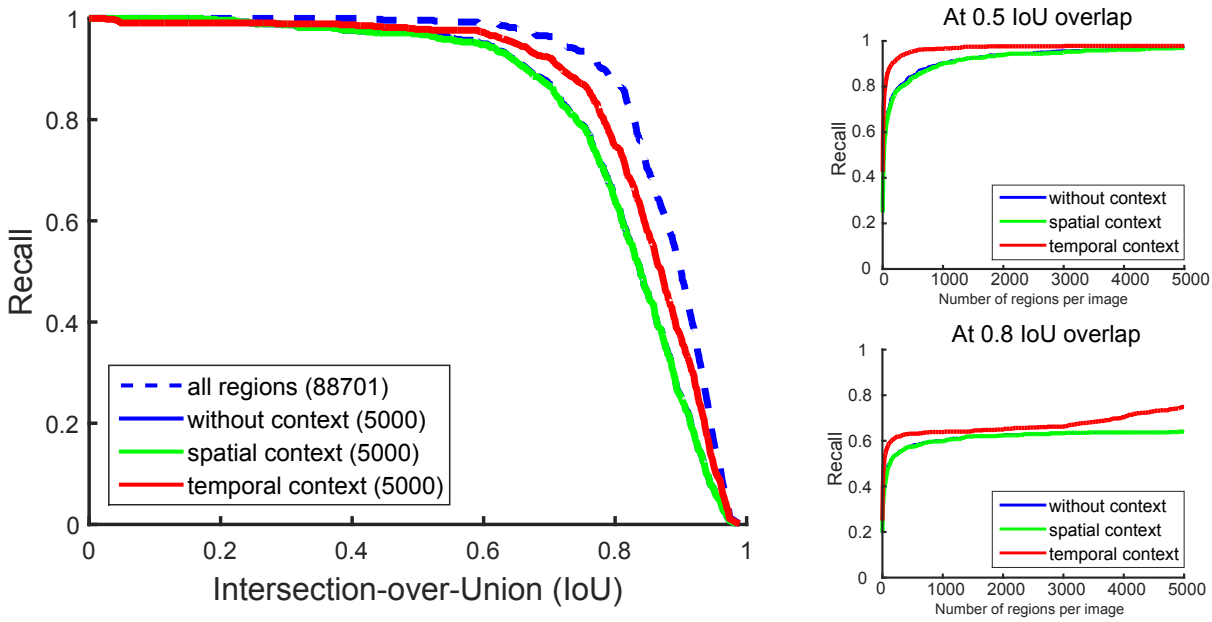


Figure 4: The recall rate evaluated on LjubljanaDFG2015 dataset. The graph on the left depicts the recall with respect to the allowed intersection-over-union with the ground-truth annotations. The dashed blue line represents the best case with the current region proposals algorithm, when using all 100.000 returned regions per image. The average number of the top N selected regions is shown in parenthesis in the legend. Note that the graph lines for spatial context and baseline without context are overlapping. The right graphs show the recall rate evaluated with the changing N, the number of regions allowed per image. The top-right graph shows only regions with an overlap of at least 0.5, and the bottom-right graph shows regions with an overlap of at least 0.8.

4 Conclusion

In this work, we proposed to integrate the non-visual context information into visual traffic sign detection in order to improve the recall performance of the region proposals. We applied the context to a two-stage detector from [1]. In particular, we integrated the additional context information into the detector's first stage to improve the ranking of region proposals. We formulated the context as a likelihood map, and constructed a context score to re-score the regions returned by the region proposals algorithm. We have experimented with two types of context information applicable to this domain. We utilized spatial context by relying on the frequency of traffic sign positions in images; however, we found no improvement when using such spatial context in region proposals. Additionally, we proposed to utilize any prior information about the 3D world coordinates of the known traffic sign positions. This information may be obtained from road maintenance authorities or from previous runs along the same route. We termed this prior information the temporal context, and showed that despite inaccuracies in the 3D world locations and noisy 3D world to 2D image projections, we are able to significantly increase the recall rate of region proposals.

In future work, we will consider integrating the spatial context into the second stage instead of the first stage. The spatial context may be more informative after classification removes the background regions that should not interfere with the re-scoring from the spatial context. We will also explore the spatial prior on the per-class basis, which may prove more accurate for traffic signs that consistently appear in the same positions.

Acknowledgment

This work was supported by ARRS research programme P2-0214, and ARRS research project L2-6765. We also thank the DFG team, in particular Simon Jud and Samo Kumar, for providing all datasets used in this work.

References

[1] D. Tabernik, R. Mandeljc, D. Skočaj, and M. Kristan, "Domain-specific adaptations for region proposals," in *CVWW*, 2015.

[2] D. Tabernik, R. Mandeljc, and D. Skočaj, "Quality of region proposals in traffic sign detection and recognition," in *ERK*, 2015, pp. 7–10.

[3] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE PAMI*, vol. 36, no. 8, pp. 1532–1545, 2014.

[4] P. Dollár and C. Zitnick, "Structured forests for fast edge detection," in *ICCV*, 2013, pp. 1841–1848.

[5] M. Kristan, "Multivariate Online Kernel Density Estimation," *Pattern Recognition*, vol. 44, pp. 2630–2642, 2011.

[6] M. Everingham, L. Van˜Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.